Fairness and Privacy Guarantees in Federated Contextual Bandits

Sambav Solanki SAMBHAV.SOLANKI@RESEARCH.IIIT.AC.IN
IIIT, Hyderabad

Shweta Jain SHWETAJAIN@IITRPR.AC.IN
IIT, Ropar

Sujit Gujar SUJIT.GUJAR@IIIT.AC.IN
IIIT, Hyderabad

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

This paper considers the contextual multi-armed bandit (CMAB) problem with fairness and privacy guarantees in a federated environment. We consider merit-based exposure as the desired *fair* outcome, which provides exposure to each action in proportion to the reward associated. We model the algorithm's effectiveness using fairness regret, which captures the difference between fair optimal policy and the policy output by the algorithm. Applying fair CMAB algorithm to each agent individually leads to fairness regret linear in the number of agents. We propose that collaborative – federated learning can be more effective and provide the algorithm Fed-FairX-LinUCB that also ensures differential privacy. The primary challenge in extending the existing privacy framework is designing the communication protocol for communicating required information across agents. A naive protocol can either lead to weaker privacy guarantees or higher regret. We design a novel communication protocol that allows for (i) Sub-linear theoretical bounds on fairness regret for Fed-FairX-LinUCB and comparable bounds for the private counterpart, Priv-FairX-LinUCB (relative to single-agent learning), (ii) Effective use of privacy budget in Priv-FairX-LinUCB. We demonstrate the efficacy of our proposed algorithm with extensive simulations-based experiments. We show that both Fed-FairX-LinUCB and Priv-FairX-LinUCB achieve near-optimal fairness regret.

Keywords: Multi-armed Bandits; Fairness; Differential Privacy; Federated Learning

1. Introduction

The *bandit* problem Auer et al. (2002) is a well-known problem encapsulating the exploration and exploitation trade-off in online learning. It has a wide array of applications, such as crowdsourcing Tran-Thanh et al. (2014), recommendation systems Li et al. (2010), sponsored search auctions Abhishek et al. (2020), service procurement Badanidiyuru et al. (2013), etc. This paper considers the contextual *multi-armed bandit* (MAB) problems in a *federated* setting.

Linear contextual bandits Li et al. (2010) associate dynamic contexts with each action by assuming that the reward for each action is modeled as a fixed but unknown linear combination of the context and thus aims to learn these linear weights for maximizing the reward of a single learning agent. Multiple agents can collaborate in many real-world applications such as crowdsourcing, service procurement, and recommender systems for better effective learning He et al. (2022); Réda et al. (2022); Solanki et al. (2022). For example, in crowdsourcing, requesters (agents) of similar tasks intend to learn the qualities of a pool of workers (actions), which are context-dependent. In such examples, agents can help each other by collaborating to learn the correlation between worker attributes (action context) and task completion proficiency (rewards) faster.

Such collaborative learning should be allowed without sharing sensitive data (such as specific worker selection in any given round) among the agents while allowing for effective learning, i.e., it should protect the privacy of individual agents' sensitive information. The literature model collaborative learning with privacy requirements via the paradigm of *federated learning* for practical collaboration Kairouz et al. (2021). Recent works Solanki et al. (2022); Dubey and Pentland (2020) have explored *differential privacy* guarantees in *federated bandits* which extend bandit problem in federated settings.

In many practical applications, actions often involve interactions with humans, e.g., workers in crowdsourcing. Here, it becomes crucial to ensure that each action receives sufficient exposure. Traditional bandit approaches exhibit a "winner takes all" behaviour Wang et al. (2021), which consistently favors the optimal action and deprives other actions of opportunity, leading to starvation among actions. We address this issue by considering *fairness of exposure* Wang et al. (2021) in multi-agent contextual bandit problems. Other fairness notions in the context of bandit problems, such as guaranteeing minimum exposure to each action Patil et al. (2020), group fairness, and fair treatment Joseph et al. (2016b) depend solely on the rewards or prioritize fairness for the learning agents rather than the individual actions. On the other hand, fairness of exposure ensures proportionality Aumann and Maschler (1985); Suksompong (2016) for the actions, meaning that every action would be selected proportional to its merit/reward. This is an essential indicator of individual fairness in ML algorithms and proportionality in game theoretical frameworks. The algorithm in Wang et al. (2021) works only for a single-agent setting. There are a few works Hossain et al. (2021); Biswas et al. (2023) that provide fairness guarantees in a federated setting; however, these works consider fairness for agents instead of actions.

Fairness of exposure in bandits focuses on minimizing *fairness regret*, which measures the deviation of action selection policy from the optimal policy satisfying fairness. For the first time, this paper provides fairness regret guarantees in the federated setting while ensuring privacy. One naive way to ensure fairness in federated learning is to adapt existing single-agent bandit techniques (e.g., from Wang et al. (2021)) by integrating a *communication protocol*. Following a naive communication protocol, all agents will communicate with each other in every round by sharing all their information about the actions. It can be easily proved the naive communication protocol, as expected, offers better fairness guarantees than the single-agent setting. However, it leads to maximum privacy leakage. Another extreme is not to allow any communication among agents. It leads to maximum privacy, but in the absence of collaborative learning, the regret blooms in terms of the number of agents. If we adopt existing communication protocols designed for privacy preservation in federated bandits, we observe that they fail to offer good regret guarantees. In summary, developing an intelligent communication protocol that provides a regret bound that is sub-linear in the number of agents and extends to the private setting is essential.

This work designs a novel communication protocol for federated bandits while learning generalizes the techniques from FairX-LinUCB Wang et al. (2021), an algorithm designed for a single-agent setting, to a federated setting. We call our algorithm as Fed-FairX-LinUCB. Our communication protocol is scalable to differentially private methods since the number of communication rounds is bounded while ensuring fairness given the bounded communication gaps. We denote the privacy-ensuring version of the proposed algorithm by Priv-FairX-LinUCB. In summary, our paper solves the fair federated contextual MAB problem while ensuring differential privacy guarantees. Our contributions include:

1. We introduce the notion of fairness for actions in federated contextual bandits.

2. We propose a novel communication protocol and show that Fed-FairX-LinUCB achieves sub-linear fairness regret in terms of the number of learning agents while being optimal in terms of the number of rounds up to a log dependence term (Theorem 6).[1]

3. The proposed communication is extensible to privatizer routine from Dubey and Pentland (2020). It lets us develop Priv-FairX-LinUCB, which ensures differential privacy guarantees (for the agents).

4. We theoretically show that Priv-FairX-LinUCB achieves differential privacy guarantees while having bounded fairness regret (Theorem 8).

5. We empirically show that Fed-FairX-LinUCB and Priv-FairX-LinUCB outperform a non-collaborative learner.

2. Related Work

Federated Bandits. Bandit problems Robbins (1952); Lai and Robbins (1985); Auer et al. (2002) with contextual nature Li et al. (2010); Abbasi-yadkori et al. (2011) have gained significant prominence in both academia and industry. Moreover, analysing bandit problems in a federated setting He et al. (2022); Réda et al. (2022) has been an important exploration of cooperative learning.

Privacy. Our work leverages federated learning, which does large dataset querying. We use differential privacy, introduced by Dwork et al. (2006), to provide privacy for context/reward information. Differential privacy is a rigorous mathematical notion of privacy that encapsulates the requirement that the probability of output should have minimal changes for neighbouring input datasets. Chan et al. (2011) and Dwork et al. (2010) introduced the notion of differential privacy under continual observation using a **tree-based algorithm**, which we leverage. This method has seen utilisation across several online learning problems Tossou and Dimitrakakis (2016); Guha Thakurta and Smith (2013); Kairouz et al. (2021); Jain et al. (2012). Shariff and Sheffet (2018) study differential privacy for the traditional contextual bandit setting but is limited to a single learning agent. Differential private federated bandits have been studied in Liu et al. (2022) and Solanki et al. (2022). However, our work is closely related to the important work of Dubey and Pentland (2020), extending it for non-traditional bandit optimisation.

Fairness in Bandits. Significant progress has been made in traditional bandits, but bandits with fairness objectives have only recently gained popularity. Joseph et al. (2016b), propose bandit fairness which is achieved by ensuring that a better arm is always chosen

[1] It is trivially implied that fairness regret would scale linearly in m for non-collaborative learning.

with at least the same likelihood as a worse arm. Several other works, including Chen et al. (2020); Patil et al. (2020), aim to guarantee a minimum exposure for arms in the stochastic bandit problem. However, based on discussion in Section 1, it remains unclear how much exposure would be enough. Hossain et al. (2021) and Biswas et al. (2023) define fairness for multi-agent setting, but fairness with respect to agents rather than actions is considered.

The notion of fairness, for actions, in the aforementioned works is modelled as a constraint rather than a desired outcome, with reward maximisation being the primary objective. In our work, we use the concept of fairness of exposure, introduced by Wang et al. (2021) for the single-agent setting and later used in works like Sood et al. (2024) and Pokhriyal et al. (2024), which is an objective-oriented notion of fairness that addresses the problem of starvation among actions. Additionally, it is important to highlight that no work has previously studied proportionality-based fairness in a federated bandit setting with respect to the actions. To the best of our knowledge, our work is the premiere work to generalize fair contextual bandits into a federated setting, in addition to being the first work to simultaneously incorporate the notion of fairness and privacy for the bandit problem.

3. Model Preliminaries

3.1. Setting and Notations

We abstract the problem as a federated contextual bandit setting where each of $M = [m]$ agents are learning about actions $a \in \mathcal{D}$. The bandit algorithm runs for T rounds, where, at each round t, an agent $i \in M$ observes a context vector $\mathcal{X}_t^i = (x_t^i(a))_{a\in\mathcal{D}}$ $(|| \, x_t^i(a) \, ||_2 \leq 1; \forall a)$ with $x_t^i(a) \in \mathbb{R}^d$ and selects an action a_t^i. Each agent observes a different context vector and selects an action independently at each round t. The agent obtains a reward for a selected action a_t^i at time t which we represent as $y_t^i(a_t^i) = \theta^* \cdot x_t^i(a_t^i) + \eta_t(a_t^i)$. Here, $\theta^* \in \mathbb{R}^d$ is an unknown but fixed parameter. As standard in the literature, $\eta_t(a_t^i)$ is a noise parameter, which is i.i.d. sub-Gaussian with mean 0. Thus, the expected reward for an action a at time t, for an agent i, is given by $\mathbb{E}[y_t^i(a)] = \theta^* \cdot x_t^i(a)$. We denote this reward by the quantity $\mu_a \mid \mathcal{X}_t^i$ representing the expected reward for an action a, when i^{th} agent is observing the context vector $\mathcal{X}_t^i$. Note that θ^* (the true parameter) is the same for all the agents and is learned by the agents till time T in a collaborative fashion while preserving the privacy of their contexts/reward observations and satisfying the fairness guarantees.

We denote the set of available contexts to all the agents at time t as $\mathcal{X}_t = (\mathcal{X}_t^i)_{i\in M}$, $\mathcal{X}^i = (\mathcal{X}_t^i)_{t=1}^{t=T}$ and $\mathcal{X} = \{\mathcal{X}^1, \mathcal{X}^2, \ldots, \mathcal{X}^m\}$. The goal of each agent i is to implement a policy $\pi_t^i(\mathcal{X}_t^i)$ which denotes the vector of probabilities of action selection by i^{th} agent at time t. The probability of selecting action a is denoted by $\pi_t^i(a, \mathcal{X}_t^i)$. Instead of maximizing the reward, each agent needs to ensure fairness amongst the actions so that all actions get a fair fraction of chances to avoid otherwise observed “winner takes it all” Mehrotra et al. (2018) problem. Specifically, this setting aims to learn a policy that selects actions with probabilities proportional to their merit. Note that the objective here is to learn the fair policy rather than the optimal-reward policy.

Agents assign a merit score function f^i over the actions based on their expected rewards for the given context. $f^i : \mathbb{R}^+ \to \mathbb{R}^+$ where $f^i(\mu_a \mid \mathcal{X}_t^i)$ denotes the score assigned by agent i for the action a when observed context is $\mathcal{X}_t^i$. Each agent then needs to implement the policy such that the following fairness constraint, which is denoted as fairness of exposure, is satisfied:

$$\frac{\pi_t^i(a, \mathcal{X}_t^i)}{f^i(\mu_a \mid \mathcal{X}_t^i)} = \frac{\pi_t^i(a', \mathcal{X}_t^i)}{f^i(\mu_{a'} \mid \mathcal{X}_t^i)} \ \forall a, a' \in \mathcal{D} \tag{1}$$

f^i quantifies the utility of rewards derived from an arm for the agent. We assume Minimum merit and Lipschitz continuity properties on merit function Wang et al. (2021). The minimum merit property provides a lower bound on the merit function, i.e. $\min_\mu f^i(\mu) \geq \gamma$, $\forall i \in M$ for some $\gamma > 0$. Lipschitz continuity property assumes that the merit function is Lipschitz continuous, i.e., $\forall \mu_1, \mu_2, i \in M$, $|f^i(\mu_1) - f^i(\mu_2)| \leq L|\mu_1 - \mu_2|$ for some $L > 0$.

We denote the optimal policy by $\pi_*^i(\mathcal{X}_t^i)$ when θ^* is known, i.e., at round t, it satisfies fairness condition (Eq. 1). Note that given a context vector the optimal policy, $\pi_*^i(.)$, does not depend on round t, $\pi_*^i(a, \mathcal{X}_t^i) = \frac{f^i(\theta^* \cdot x_t^i(a))}{\sum_{a' \in \mathcal{D}} f^i(\theta^* \cdot x_t^i(a'))}$. Typically, θ^* being unknown, each agent is learning θ^* and in turn the optimal policy through algorithm $\mathcal{A}$ over the rounds, taking actions using policy $\pi_t^i(\cdot)$. $\hat{\theta}_t^i$ is used to denote the learnt θ^* for agent i at time t. Unlike the optimal policy, $\pi_t^i(\cdot)$ is round dependent. For agent i at round t the *instantaneous fairness regret* is defined as: $FR_t^i(\mathcal{A}, \mathcal{X}_t^i) = \sum_{a \in \mathcal{D}} |\pi_*^i(a, \mathcal{X}_t^i) - \pi_t^i(a, \mathcal{X}_t^i)|$. As these agents learn about the same actions, they can communicate with each other about their estimates of θ^* and learn it faster, reducing the per-agent fairness regret. We assume that all the agents deploy the same learning algorithm. Thus, the *fairness regret* can be defined as:

Definition 1 *Fairness Regret. For a learning algorithm* $\mathcal{A}$, *we define fairness regret as* $FR(\mathcal{A}, T, \mathcal{X}) = \frac{1}{m} \sum_{i \in M} FR^i(\mathcal{A}, T, \mathcal{X}^i)$ *where* $FR^i(\mathcal{A}, T, \mathcal{X}^i) = \sum_{t=1}^T FR_t^i(\mathcal{A}, \mathcal{X}_t^i)$

Henceforth, we will avoid using $\mathcal{X}_t^i$ from fairness regret to avoid notation clutter. Additionally, since we are bounding it only for the algorithms in the paper, we refer to the above quantities as FR_t^i, FR_i, FR. We also use $FR^i([T_1, T_2])$ to denote $\sum_{t=T_1}^{t=T_2} FR_t^i$ and similarly $FR([T_1, T_2])$.

3.2. Why fairness of exposure?

We motivate with a single agent setting who is interested in assigning tasks to 3 workers with unknown completion times. Let the optimal task assignment (according to Eq. 1)distribution be $[0.14, 0.28, 0.56]$, where faster worker is assigned more tasks, if the goal is to minimize total project completion time while ensuring exposure guarantees to the workers. Traditional regret optimization finds the best worker which does not lead to balanced/fairer task allocation.While some approaches try to incorporate fairness into bandit algorithms, they often fall short in the task assignment scenario:

- Delta-fairness Joseph et al. (2016a); R and Dukkipati (2020), which prioritizes arms (workers) with higher rewards will essentially lead to giving maximum tasks to optimal (faster) worker, in this case the worker 3, however it does not provide any exposure guarantee.

- Minimum share fairness Patil et al. (2020); Chen et al. (2020)ensures each worker receives a minimum fraction of tasks. Utility optimisation in this case relies on knowing expected completion times, which are unknown in our problem. This makes its effectiveness uncertain.

In contrast, proportionality-based fairness offers a more promising approach by directly aligning fairness with utility optimization. Furthermore, when workers are involved in multiple projects simultaneously, (i.e., multiple agents are learning about the workers) federated learning with differential privacy can further optimize task assignment by sharing limited information privately, leading to faster learning and improved project completion times.

3.3. Fairness in Single-Agent Contextual MAB

We start with some notation and summarize FairX-LinUCB for a single-agent MAB setting Wang et al. (2021).

- If H is positive semi-definite matrix, it is represented by $H \succeq 0$. Additionally, for two matrices H_1 and H_2, $H_1 \succeq H_2$ implies $H_1 - H_2 \succeq 0$.
- The $H - norm$ for vector y w.r.t. a positive semi-definite matrix H, is denoted by $\|y\|_H = \sqrt{y^\intercal H y}$

The central idea is to construct a confidence region, CR_t, at every round t, containing θ^* with high probability. The confidence region is an ellipsoid centered around the linear regression estimate $\hat{\theta}_t = (I\lambda + X_{<t}X_{<t}{}^\intercal)^{-1}X_{<t}{}^\intercal Y_{<t}$. Here, $X_{<t} = [x_1(a_1)^\intercal \dots x_{t-1}(a_{t-1})^\intercal]^\intercal$, $Y_{<t} = [y_1(a_1) \dots y_{t-1} \ (a_{t-1})]^\intercal$, and $x_t(a_t)$ and $y_t(a_t)$ denotes the context and observed reward of the selected action a_t at time t respectively. The proposed algorithm then optimistically selects θ_t from the confidence region, and the selection policy, π_t, using θ_t based on the constraints. The selection policy defines a probability distribution over the actions, based on which an action is chosen, and the observed rewards for the chosen action are used to improve the estimation further. Optimistic selection is a non-convex-constrained optimization problem, and projected gradient descent is used to find approximate solutions.

3.4. Privacy requirements

We consider privacy over the agent-action interaction, i.e., for any agent i, we consider that the context vectors $(\mathcal{X}^i)$ and the observed feedback $((y_t^i(a_t^i))_{t\in[T]})$ should be kept private. Considering that agent only needs to store $x_t^i(a_t^i)$ for feedback estimation, we use the differential privacy definition with respect $(x_t^i(a_t^i), y_t^i(a_t^i))_{t\in[T]}$. Our differential privacy notion matches the one defined in Dubey and Pentland (2020). Here, we leverage their differential privacy definition for our setting. Let us consider two sets $\mathcal{S}_i = (x_t^i(a_t^i), y_t^i(a_t^i))_{t\in[T]}$ and $\mathcal{S}_i{}' = (x_t^i(a_t^i)', y_t^i(a_t^i)')_{t\in[T]}$. They are considered to be $t' - neighbors$ if at all time steps $t \neq t'$, $(x_t^i(a_t^i), y_t^i(a_t^i)) = (x_t^i(a_t^i)', y_t^i(a_t^i)')$.

Definition 2 *Federated Differential Privacy (Dubey and Pentland, 2020, Definition 1) In a federated learning setting with $m \geq 2$ agents, a randomized multi-agent contextual bandit algorithm $\mathcal{A} = (\mathcal{A}^i)_{i=1}^m$ is $(\epsilon, \delta, m)-$ federated differentially private under continual multi-agent observation if for any $i, j \in M$ such that $i \neq j$, any t and set of sequences $\mathbb{S}_i = (\mathcal{S}_k)_{k=1}^m$ and $\mathbb{S}_i{}' = (\mathcal{S}_k)_{k=1,k\neq i}^m \bigcup \mathcal{S}_i{}'$ such that $\mathcal{S}_i{}'$ and $\mathcal{S}_i$ are $t'-$neighbors, and any subset of actions $(a_t^j)_{t\in[T]} \subset \mathcal{D} \times \ldots \times \mathcal{D}$ of actions, it holds that:*

$$\mathbb{P}(\mathcal{A}^j(\mathbb{S}_i) \in (a_t^j)_{t\in[T]}) \leq e^\epsilon \mathbb{P}(\mathcal{A}^j(\mathbb{S}_i{}') \in (a_t^j)_{t\in[T]}) + \delta$$

Here, the quantity, $\mathcal{L}^{o}_{A^j(\mathbb{S}_i)||A^j(\mathbb{S}'_i)} = \log(\frac{\mathbb{P}(\mathcal{A}^j(\mathbb{S}_i)\in o)}{\mathbb{P}(\mathcal{A}^j(\mathbb{S}_i{}')\in o)})$ refers to privacy loss incurred by observing output $o = (a^j_t)_{t\in[T]}$.

Differential privacy ensures that the presence/absence of one data point does not lead to significant learning changes – for federated bandit settings, it implies small changes in the data sharing do not lead to any major action changes.

Goal: Each agent's goal is to learn θ^* while minimizing fairness regret (Definition 1); albeit ensuring differential privacy guarantees (Definition 2).

4. Multi-Agent Fair and Private Contextual Bandit Algorithm

The communication protocol currently used in federated bandits literature is not suitable for achieving bounded fairness regret. It is important to limit the number of communication rounds and maintain a constrained gap between communication instances in order to ensure both bounded fairness regret, and scalability with private methods. The total privacy loss, which is the composition of privacy losses incurred overall communication rounds, is proportional to the number of communication rounds. Thus, it follows that for a budgeted (fixed) total privacy loss, the maximum possible per-round privacy loss is inversely proportional to the number of communication rounds. As a result, the number of communication rounds should be bounded to control the accumulation of noise and maintain privacy within acceptable limits. At the same time, bounding the gaps between communication rounds is necessary to make fairness regret claims.

In this section, we firstly build an algorithm, Fed-FairX-LinUCB, that learns θ^* collectively amongst m agents using a novel communication protocol. We then design a privacy-preserving version, Priv-FairX-LinUCB, in Section 4.2.

4.1. Fed-FairX-LinUCB

We consider a group of m agents actively participating in the contextual bandit problem and maintaining synchronization through periodic communication. Algorithm 1 without the privatizer routine represents Fed-FairX-LinUCB. Essentially, the exact information of the agents is sent to other agents when communication is required. For any agent i, at round t, let the last synchronization round take place at instant t'. Then, there exist two sets of parameters. The first set of parameters is the set of all observations made by all m agents till round t'. We store this in terms of a shared gram matrix, $U_t = \sum_{i\in M}(\lambda I + \sum_{\tau=1}^{t'}(x^i_\tau(a^i_\tau))(x^i_\tau(a^i_\tau))^\intercal)$, and a shared vector, $u_t = \sum_{i\in M}\sum_{\tau=1}^{t'}(x^i_\tau(a^i_\tau))y^i_\tau(a^i_\tau)$. Secondly, each agent has access to its own observations since the last communication round. We note those using the gram matrix $S^i_t = \sum_{\tau=t'}^{t}(x^i_\tau(a^i_\tau))(x^i_\tau(a^i_\tau))^\intercal$ and the reward vector $s^i_t = \sum_{\tau=t'}^{t}(x^i_\tau(a^i_\tau))y^i_\tau(a^i_\tau)$, where t' was the last communication round. The agents use combined parameters for estimating the linear regression estimate, $\hat{\theta}^i_t$. For an agent i, $V^i_t = U_t + S^i_t$, $b^i_t = u_t + s^i_t, \hat{\theta}^i_t = (V^i_t)^{-1}b^i_t$. The agents then constructs a confidence region, CR^i_t around $\hat{\theta}^i_t$. Suitable sequence $[\sqrt{\beta^i_t}]_{i\in M, t\in[T]}$ needs to be used, ensuring that with high probability $\forall i, t,\ \theta^* \in CR^i_t$. An optimistic estimate, θ^i_t is selected from CR^i_t (line 6 of Algorithm. 1). The agent selects the action using a policy construction, π^i_t. This ensures fairness by assigning a probability distribution for action selection based on estimated merit. We now explain our communication protocol that achieve sub-linear fairness regret.

Algorithm 1 Priv-FairX-LinUCB

```
1:  Input: $\beta_t$, $[f^i]_{\forall i \in [m]}$, $\lambda$, $m$
2:  Initialization: $\forall i \in [m]$, $V_1^i = S_1^i = U_1 = \lambda \mathbf{I}_d$, $b_1^i = s_1^i = u_1 = \mathbf{0}_d$, $\tau = 1$.
3:  for $t = 1$ to $T$ do
4:    for $i = 1$ to $m$ do
5:      Observe contexts $\mathcal{X}_t^i$; $\hat{\theta}_t^i = (V_t^i)^{-1} b_t^i$; $\mathbf{CR}_t^i = (\theta : \|\theta - (\hat{\theta}_t^i)\|_{V_t^i} \leq \sqrt{\beta_t^i})$
6:      $\theta_t^i = \text{argmax}_{\theta \in CR_t^i} \sum_{a \in \mathcal{D}} \frac{f(\theta x_t^i(a))}{\sum_{a' \in \mathcal{D}} f(\theta x_t^i(a'))} \theta x_t^i(a)$
7:      Construct Policy $\pi_t^i(a) = \frac{f(\theta_t^i x_t^i(a))}{\sum_{a'} f(\theta_t^i x_t^i(a'))}$
8:      Sample arm $a_t^i \sim \pi_t^i$ and observe reward $y_t^i(a_t^i)$
9:      $S_{t+1}^i = S_t^i + (x_t^i(a_t^i))(x_t^i(a_t^i))^\intercal$; $s_{t+1}^i = s_t^i + (x_t^i(a_t^i)) y_t^i(a_t^i)$
10:     if $t == \tau$ then
11:       $Sync \longleftarrow true$
12:       if $t < \lceil \frac{T}{md^2 \log^2(1+T/d)} \rceil$ then
13:         $\tau = 2\tau$
14:       else
15:         $\tau = \tau + \lceil \frac{T}{md^2 \log^2(1+T/d)} \rceil$
16:       end if
17:     end if
18:     if $Sync$ then
19:       $[\forall j \in M]$ Send $S_t^j, s_t^j \rightarrow PRIVATIZER$
20:       $[\forall j \in M]$ Receive $\hat{U}_t^j, \hat{u}_t^j \leftarrow PRIVATIZER$
21:       $[\forall j \in M]$ Communicate $\hat{U}_t^j, \hat{u}_t^j$ to others
22:       $[\forall j \in M]$ $U_{t+1} = \sum_{k=1}^{M} \hat{U}_t^k$; $u_{t+1} = \sum_{k=1}^{M} \hat{u}_t^k$; $S_t^j = \mathbf{0}_{d \times d}$; $s_t^j = \mathbf{0}_d$; $\Delta_t^j = 0$
23:       $Sync \longleftarrow false$
24:     else
25:       $U_{t+1} = U_t^i$; $u_{t+1} = u_t^i$; $\Delta_{t+1}^i = \Delta_t^i + 1$
26:     end if
27:     $V_{t+1}^i = U_{t+1} + S_{t+1}^i$; $b_{t+1}^i = u_{t+1} + s_{t+1}^i$
28:   end for
29: end for
```

Communication Protocol. If the agents were to communicate in every round without any optimization, they could enhance their fairness regret by order of $O(1/\sqrt{m})$. However, communicating at every round results in inefficiencies and potential privacy breaches. To address these concerns, our algorithm suggests a communication strategy allowing agents to communicate only $\lceil 2md^2 \log^2(1 + T/d) \rceil$ times while achieving comparable fairness regret performance. In our proposed approach, we suggest that the agents communicate with increasing intervals between two consecutive communication rounds during the first $\lceil \frac{T}{md^2 \log^2(1+T/d)} \rceil$ rounds (line 12-13 of Algorithm 1). Subsequently, they communicate only after every $\lceil \frac{T}{md^2 \log^2(1+T/d)} \rceil$ rounds. Rapid communication in the initial rounds proves beneficial in practice, considering the trend in regret is sub-linear in T. Concurrently, the number of communication rounds and the gap between the communication rounds remain bounded. The number of communication rounds is upper bounded by $O\left(\log(\lceil\frac{T}{md^2\log^2(1+T/d))}\rceil) + md^2\log^2(1+T/d)\right)$ while the gap between communication rounds is upper bounded by $\lceil\frac{T}{md^2\log^2(1+T/d)}\rceil$, both of which can be trivially calculated. This distinguishes it from the communication protocols employed by Dubey and Pentland (2020); Solanki et al. (2022), where the gaps between communication rounds can be of the order $O(T)$, which makes it difficult to bound fairness regret. In summary, on observing the context set, each agent utilizes their estimate of θ^* to formulate a selection policy, which yields a probability distribution for choosing an action. Once an action is selected and the corresponding reward is observed, the agents update their local estimates and periodically exchange these updates with each other to enhance the accuracy of the shared estimates.

4.2. Priv-FairX-LinUCB

The key difference between Priv-FairX-LinUCB and Fed-FairX-LinUCB lies in the communication perturbation. In a non-private setting, we communicate exact observations about context and reward to all other agents. However, we must carefully add perturbation for the private setting to satisfy the differential privacy constraints mentioned in section 3. In the private setting, let $\hat{U}_t^i = \sum_{\tau=1}^{t-1}(x_\tau^i(a_\tau^i))(x_\tau^i(a_\tau^i))^\intercal + H_t^i$, $\hat{u}_t^i = \sum_{\tau=1}^{t-1}(x_\tau^i(a_\tau^i))y_\tau^i(a_\tau^i) + h_t^i$ denote the perturbed contexts and rewards. Here H_t^i and h_t^i are noise additions used for perturbation. Here, $V_t^i = \sum_{i\in M}\hat{U}_t^i + S_t^i$ and $b_t^i = \sum_{i\in M}\hat{u}_t^i + s_t^i$, where S_t^i and s_t^i remains same as stated in Section 4.1. We note that V_t^i can also be represented as: $V_t^i = G_t^i + H_t^i$ with G_t^i denoting the gram matrix in absence of noise perturbations.

To achieve privacy, we introduce a privatized version of the synchronization process amongst the agents. We do so by using the privatizer routine, which uses a tree-based mechanism to communicate while limiting the noise addition. The tree-based mechanism for differential privacy maintains a binary tree of logarithmic depth in terms of communication rounds. The sequential data released at communication rounds are stored at the leaf nodes, while every parent node stores the sum of the child nodes' data. In addition, noise is sampled at each node to maintain privacy. This allows for returning partial sums by adding at max k nodes if k was the depth of the tree. While our algorithm vastly differs from the FedUCB algorithm Dubey and Pentland (2020) in terms of objective constraint, arm selection protocol, and communication round selection, it resembles our algorithm in terms of linear regressor estimation in a federated setting. Based on this, we can use the privatizer routine with marginal changes to ensure privacy guarantees. The privatizer routine is formally outlined for completeness.

5. Theoretical Analysis

On a high level, the fairness regret proof considers a single hypothetical agent who plays mT rounds instead of considering m agents playing T rounds, each with sparse communication. The bounded deviation from this scenario to our intended setting is used to show the fairness regret analysis. Lemma 3 captures the fairness regret in terms of the determinant of the gram matrices, which is important to capture the deviation between the hypothetical agent and our intended set of agents, while lemma 4 is useful for fairness regret bounds

Algorithm 2 PRIVATIZER

```
Input: $\epsilon, \delta, d, \tau$ (number of communication rounds), $L$ (upper bound on norm of context vector)
Initialization:
$n = 1 + \lceil \log \tau \rceil$
$\mathcal{T} \leftarrow$ a binary tree of depth $n$
for each node $i$ in $\mathcal{T}$ do
    Create a noise matrix: $\hat{N} \in \mathbb{R}^{d \times (d+1)}$, where $\hat{N}_{kl} \sim \mathcal{N}(0, 16n(L^2 + 1)^2 \log(2/\delta)^2/\epsilon^2)$
    $N = (\hat{N} + \hat{N}^\intercal)/\sqrt{2}$
end for
Runtime:
for each communication round $t$ do
    Receive $S_t^i, s_t^i$ from agent, and insert it into $\mathcal{T}$ as a $d \times (d+1)$ matrix (Alg. 5, Jain et al. (2012))
    Receive $M_t^i$ using the least nodes of $\mathcal{T}$ (Alg. 5, Jain et al. (2012))
    $\hat{U}_t^i = U_t^i + H_t^i$, top-left $d \times d$ submatrix of $M_t^i$
    $\hat{u}_t^i = u_t^i + h_t^i$, last column of $M_t^i$
    Return $\hat{U}_t^i, \hat{u}_t^i$
end for
```

for a single-agent. Lemma 5 formalizes the instantaneous fairness regret, a prerequisite for proving Theorem 6.

5.1. Regret Analysis

The following lemma is useful in proving the fairness regret of Fed-FairX-LinUCB.

Lemma 3 *(Elliptical Potential (Shariff and Sheffet, 2018, Lemma 22)). Let $x_1, \ldots, x_n \in R^d$ be vectors with each $\|x_t\| \leq L$. Given a positive definite matrix $U_1 \in R^{d \times d}$, define $U_{t+1} := U_t + x_t x_t^\top$ for all t. Then $\sum_{t=1}^{n} \min\left\{1, \|x_t\|^2_{U_t^{-1}}\right\} \leq 2 \log \frac{\det U_{n+1}}{\det U_1} \leq 2d \log \frac{\operatorname{tr} U_1 + nL^2}{d \det^{1/d} U_1}$*

Also, we extend Lemma A.6.4 from Wang et al. (2021) to multi-agent setting as follows.

Lemma 4 *When $\| \; x_t^i(a) \; \|_2 \leq 1 \; \forall a, t, i$, for the Fed-FairX-LinUCB algorithm, $\forall i \in [m]$, with probability $1 - \delta/2$,*

$$\left| \sum_{t=1}^{T} w_t^i(a_t^i) - \sum_{t=1}^{T} \mathbb{E}_{a \sim \pi_t^i} w_t^i(a) \right| \leq \sqrt{2T \ln(4/\delta)}$$

Here, $w_t^i(a) = \sqrt{x_t^i(a)(V_t^i)^{-1}(x_t^i(a))^\intercal}$ is the normalized width. With the help of the above lemmas, we now provide bounds on instantaneous regret FR_t^i and defer proofs to appendix.

Lemma 5 *For the Fed-FairX-LinUCB, with high probability, the instantaneous regret for any agent i is bounded by,*

$$FR_t^i = \sum_{a \in \mathcal{D}} \left| \pi_t^i - \pi_*^i \right| \leq \frac{4L\sqrt{\beta_t}}{\gamma} \mathbb{E}_{a \sim \pi_t^i} \left\| x_t^i(a) \right\|_{(V_t^i)^{-1}}$$

The probability with which Lemma 5 holds true is dependent on β_t^i, where $\beta_t = max_{i \in M} \beta_t^i$.

Theorem 6 *With high probability, Fed-FairX-LinUCB achieves a fairness regret of*

$$O\left(\frac{4\nu L\sqrt{\beta_t}}{\gamma}\sqrt{mTd\log{(1+\frac{T}{d})}+m^2d^3\log^3{(1+\frac{T}{d})}}\right)$$

when $\| x_t^i(a) \|_2 \leq 1 \, \forall a, t, i$.

The values in sequence of β_t dictates the probability with which Lemma 5, and in turn Theorem 6 holds. The problem of selection of values in sequence of β_t is well studied in the literature. For instance, using Theorem 2 from Abbasi-yadkori et al. (2011), it can be said that θ^* lies in the confidence region with probability $1-\alpha$ for $\beta_t = O\left(d\log{(\frac{1+mt}{\alpha})}\right)$ resulting in a regret bounds of $\tilde{O}\left(d\sqrt{mT\log^2{(1+mT/d)}}\right)$ for Fed-FairX-LinUCB (typically $m << T$ and hence the $\sqrt{mT}$ term dominates $\sqrt{m^2\log^2(1+T/d)}$).

The key difference between private and non-private regret analysis lies in the gram matrix regularization and confidence interval construction (use of appropriate β_t).

We note the following claim is useful for completing Priv-FairX-LinUCB's regret analysis. It provides values for the sequence of β_t for which the confidence interval contains θ^* with high probability.

Lemma 7 *(Similar to (Dubey and Pentland, 2020, Proposition 2)) For an instance of problem where synchronisation occurs exactly n times in a span of T trials, and $\underline{\rho}, \bar{\rho}$ and z are $(\alpha/2nm)$-accurate (Dubey and Pentland, 2020, Definition 3). Then for Priv-FairX-LinUCB with bounded target parameter ($\| \theta^* \|_2 \leq c$), the sequence of $\sqrt{\beta_t^i}$ is (α, M, T)-accurate if, $\sqrt{\beta_t^i} = \sigma\sqrt{2\log{(\frac{2}{\alpha})}+d\log{(\frac{\bar{\rho}}{\underline{\rho}}+\frac{t}{d\underline{\rho}})}} + mc\sqrt{\bar{\rho}} + mz$*

Theorem 8 *With high probability, when $\| x_t^i(a) \|_2 \leq 1 \, \forall a, t, i$ and Lemma 7 holds, Priv-FairX-LinUCB achieves a fairness regret of*
$O\left(\frac{4\nu L\sqrt{\beta_T}}{\gamma}\sqrt{mTd\log{(\frac{\bar{\rho}}{\underline{\rho}}+\frac{T}{d\underline{\rho}})}+m^2d^3\log^3{(\frac{\bar{\rho}}{\underline{\rho}}+\frac{T}{d\underline{\rho}})}}\right)$.

5.2. Privacy Guarantees

As mentioned in Section 4.2, we can leverage the privatizer routines to provide differential privacy guarantees for Priv-FairX-LinUCB. At each synchronization, new observations, S_t^i and s_t^i, are added to a leaf node, while all other nodes store the sum of the child nodes. Thus, $1+\lceil\log(n)\rceil$ nodes of the tree, where n is the total number of communication rounds, are sufficient to represent any partial sum till the last synchronization round. Since the privatizer routine follows the routine introduced by earlier works, it trivially follows that if each node guarantees $(\epsilon/\sqrt{8m\ln{(2/\delta)}}, \; \delta/2m)-$privacy, the outgoing communication is guaranteed to be $(\epsilon, \delta, m)-$federated differential private for each synchronization with similar values for $\bar{\rho}, \underline{\rho}, z$.

Claim 1 *(Follows from (Dubey and Pentland, 2020, Remark 3)) The privatizer routine in Priv-FairX-LinUCB guarantees that each of the outgoing messages for an agent i is $(\epsilon, \delta)-$differentially private.*

6. Experimental Analysis

6.1. Experimental Set-up

Dataset Synthetic datasets were generated for all experiments by randomly fixing the model parameter θ^*. Context size was set to five ($d = 5$), and feature vectors $\mathcal{X}_t^i$ were sampled from a uniform distribution, $x_t^i(a) \in [0,1]^d$. Noise $\eta_t^i(a)$, sampled from a normal distribution centered at 0, was added to produce reward observations.

Merit Function and Optimization A steep merit function,$f(\cdot) = e^{10\mu}$, was employed, similar to Wang et al. (2021). Projected gradient descent was used in each round to solve the resulting non-convex optimization problem.

6.2. Evaluation Set-up

Evaluation Metric Fairness regret was used as the primary evaluation metric to assess the algorithms' ability to balance performance and fairness. Exp 1 and 2 shows fairness regret trends with respect to rounds while Exp 3 and 4 uses the fairness regret at $t = 100,000$. The objective is to minimise fairness regret, and thus it is being used as the evaluation metric in the experiments. (Though our focus is on fairness, for completeness, we also evaluate the proposed algorithms for reward regret Wang et al. (2021) in Appendix.)

Experiment Repetition All reported results were averaged over 5 runs to ensure statistical significance.

6.3. Baselines

As we propose a novel setting, there are no algorithms for direct comparisons. 2 different kinds of baselines are used to demonstrate the efficacy of our proposed algorithm.

Single-Agent Baseline ($B0$) FairX-LinUCB algorithm was employed as a single-agent baseline to facilitate comparison with federated learning approaches. We note it as $B0$ in our experiments. Each agent essentially learns on their own do not communicate with other agents under this baseline.

Communication Protocol Baseline ($B1$, $B2$) Two existing communication protocols from Dubey and Pentland (2020) and Solanki et al. (2022) were compared against the proposed protocol to evaluate its efficacy. These have been termed $B1$ and $B2$ respectively. Note that the algorithms proposed in Dubey and Pentland (2020) and Solanki et al. (2022) optimize for traditional regret, hence Priv-FairX-LinUCB has been modified to just use their proposed communication protocols to form $B1$ and $B2$.

6.4. Experiments [2]

Exp 1: Single-Agent vs Federated Learning Compares the fairness regret of baseline $B0$ to the proposed non-private algorithm, Fed-FairX-LinUCB and its differentially private counterpart, Priv-FairX-LinUCB, for 10 agents (m). [$\epsilon = 2$, $\delta = 0.1$, $t \in [1, 100000]$]

[2] source code will be made publicly available

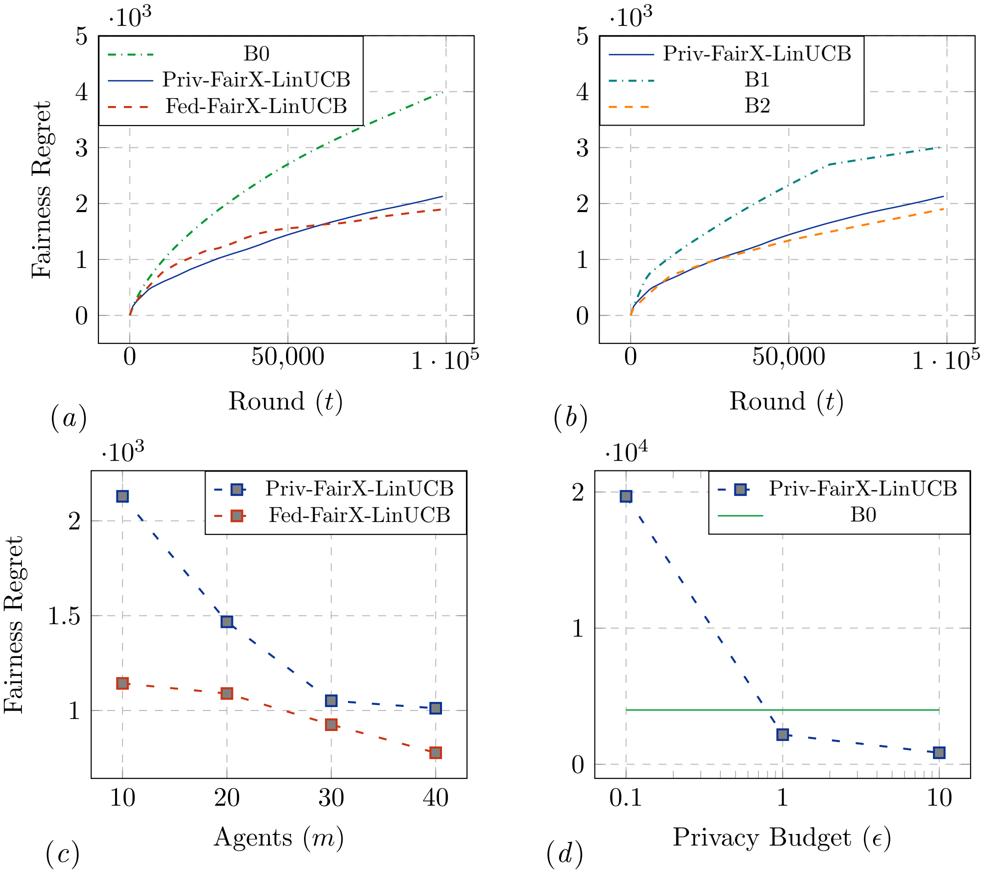

Figure 1: **(a)** Exp 1 : Fairness Regret vs. Rounds for single-agent baseline and proposed federated learning algorithms (m=10) **(b)** Exp 2 : Fairness Regret vs. Rounds for different communication protocol baselines and proposed algorithms (m=10) **(c)** Exp 3 : Fairness Regret trend w.r.t. number of agents (t=100,000) **(d)** Exp 4 : Fairness Regret trend w.r.t. privacy budget (t=100,000)

Exp 2: Communication Protocol Assesses the performance Priv-FairX-LinUCB against $B1$ and $B2$ with 10 agents. $[\epsilon = 2,\ \delta = 0.1,\ t \in [1, 100000]]$
Exp 3: Dependence on m Compares the impact of the number of agents (m) on the fairness regret of both proposed algorithms. $[\epsilon = 2,\ \delta = 0.1,\ t = 100000]]$
Exp 4: Privacy Budget Examines the effect of the privacy budget (ϵ) on the fairness regret of the private algorithm. $[m = 10,\ \delta = 0.1,\ t = 100000]]$

6.5. Inferences

- Both federated learning algorithms outperformed the single-agent baseline in terms of fairness regret.
- Priv-FairX-LinUCB outperforms B1 while producing comparable performance for B2. But unlike B2, Priv-FairX-LinUCB has bounded communication gaps, which is neces-

sary for the theoretical guarantees provided. In B2, communication gaps are as high as $O(T)$ in the later stages, and hence, in theory, fairness regrets could be as bad as $O(T)$ for B2.

- The fairness regret scales as expected with respect to the number of agents, validating theoretical results.
- The private algorithm achieved reasonable performance for ϵ values of 1 or greater, highlighting the trade-off between privacy and regret.

7. Conclusion

This work looked at the federated contextual bandit problem with fairness optimization objective while ensuring privacy of the agents. In order to extend the FairX-LinUCB algorithm to a federated setting, we proposed a novel communication protocol in Algorithm 1. Through rigorous theoretical analysis, we proved that Fed-FairX-LinUCB, the non-private algorithm achieves sub-linear fairness regret (compared to linear regret in a non-federated setting) with respect to the number of agents, i.e., for the m agent setting, the total fairness regret is of the order $O(\sqrt{m})$ (Theorem 6). Additionally, we presented Priv-FairX-LinUCB, which ensured differential privacy guarantees for the agents. We show that Priv-FairX-LinUCB has bounded fairness regret (Theorem 8). We empirically validated our results, highlighting that Priv-FairX-LinUCB significantly improves over non-collaborative learning while maintaining privacy.

We believe that alternate objectives, such as fairness, are essential for the practical adoption of the bandit framework in many use cases and that this work helps pave the way for other exciting works in that direction besides enabling suitable applications.

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

Kumar Abhishek, Shweta Jain, and Sujit Gujar. Designing truthful contextual multi-armed bandits based sponsored search auctions. *arXiv preprint arXiv:2002.11349*, 2020.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 2002.

Robert J Aumann and Michael Maschler. Game theoretic analysis of a bankruptcy problem from the talmud. *Journal of Economic Theory*, 1985.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks: Dynamic procurement for crowdsourcing. In *The 3rd Workshop on Social Computing and User Generated Content, co-located with ACM EC*, 2013.

Arpita Biswas, Jackson A Killian, Paula Rodriguez Diaz, Susobhan Ghosh, and Milind Tambe. Fairness for workers who pull the arms: An index based policy for allocation of restless bandit tasks. *arXiv preprint arXiv:2303.00799*, 2023.

T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Trans. Inf. Syst. Secur.*, 2011.

Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Abhimanyu Dubey and AlexSandy' Pentland. Differentially-private federated linear bandits. *Advances in Neural Information Processing Systems*, 2020.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference*, 2006.

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*, 2010.

Abhradeep Guha Thakurta and Adam Smith. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In *Advances in Neural Information Processing Systems*, 2013.

Jiafan He, Tianhao Wang, Yifei Min, and Quanquan Gu. A simple and provably efficient algorithm for asynchronous federated contextual linear bandits. In *Advances in Neural Information Processing Systems*, 2022.

Safwan Hossain, Evi Micha, and Nisarg Shah. Fair algorithms for multi-agent multi-armed bandits. *Advances in Neural Information Processing Systems*, 2021.

Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In *Conference on Learning Theory*, 2012.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Fair algorithms for infinite and contextual bandits. *arXiv preprint arXiv:1610.09559*, 2016a.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, 2016b.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 2021.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 2010.

Xutong Liu, Haoru Zhao, Tong Yu, Shuai Li, and John CS Lui. Federated online clustering of bandits. In *Uncertainty in Artificial Intelligence*, 2022.

Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y Narahari. Achieving fairness in the stochastic multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Subham Pokhriyal, Shweta Jain, Ganesh Ghalme, Swapnil Dhamal, and Sujit Gujar. Simultaneously achieving group exposure fairness and within-group meritocracy in stochastic bandits. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 1576–1584, 2024.

Shaarad A. R and Ambedkar Dukkipati. A regret bound for non-stationary multi-armed bandits with fairness constraints. *CoRR*, abs/2012.13380, 2020.

Clémence Réda, Sattar Vakili, and Emilie Kaufmann. Near-optimal collaborative learning in bandits. *arXiv preprint arXiv:2206.00121*, 2022.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. *Advances in Neural Information Processing Systems*, 2018.

Sambhav Solanki, Samhita Kanaparthy, Sankarshan Damle, and Sujit Gujar. Differentially private federated combinatorial bandits with constraints. *arXiv preprint arXiv:2206.13192*, 2022.

Archit Sood, Shweta Jain, and Sujit Gujar. Fairness of exposure in online restless multi-armed bandits. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2474–2476, 2024.

Warut Suksompong. Asymptotic existence of proportionally fair allocations. *Mathematical Social Sciences*, 2016.

Aristide CY Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Long Tran-Thanh, Sebastian Stein, Alex Rogers, and Nicholas R Jennings. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence*, 2014.

Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*, 2021.